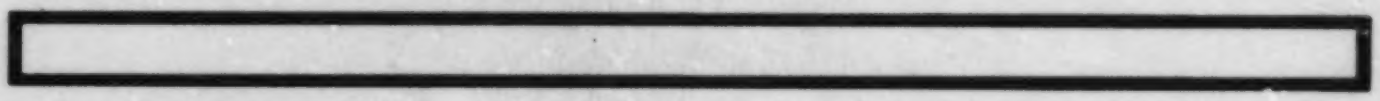
PB81-192023

Multiple Probit Analysis with a Nonzero Background

(U.S.) Health Effects Research Lab.
Research Triangle Park, NC

1980



U.S. Department of Commerce
National Technical Information Service
NTIS

PB81-192023

EPA-600/J-80-256
Journal Article

Multiple Probit Analysis with a Nonzero Background

Victor Hasselblad, Andrew G. Stead and John P. Creason

Reprinted from
BIOMETRICS [redacted] 1980
THE BIOMETRIC SOCIETY, Vol. 36, No. 4, December 1980

pp. 659-663

PRODUCT OF:
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing) MS-79-044

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA-600/J-80-256	JOURNAL ARTICLE	PB81 192023

4. TITLE AND SUBTITLE	5. REPORT DATE
Multiple Probit Analysis with a Nonzero Background	April 1981
	6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Victor Hasselblad, Andrew G. Stead and John P. Creason	

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
Biometry Division Health Effects Research Laboratory EPA-Research Triangle Park, N.C. 27711	A9GA1A 11. CONTRACT/GRANT NO.

12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Office of Research and Development Health Effects Research Laboratory EPA-Research Triangle Park, N.C. 27711	14. SPONSORING AGENCY CODE EPA-600/11

15. SUPPLEMENTARY NOTES

Published In: Biometrics 36(4):659-663, 1980

16. ABSTRACT

The 'EM' (Expectation-Maximization) algorithm is applied to probit analysis with multiple independent variables and a nonzero response rate. The equations for the maximum likelihood estimators are relatively simple, and converge in all the cases so far examined. An animal bioassay example is used to demonstrate the method. Likelihood ratio tests are used for significance tests in the analysis.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	
	20. SECURITY CLASS *(This page)* UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (9-73)

BIOMETRICS 36, 659–663
December 1980

Multiple Probit Analysis with a Nonzero Background

Victor Hasselblad, Andrew G. Stead and John P. Creason

Biometry Division, Health Effects Research Laboratory, Environmental Research Center, Environmental Protection Agency, Research Triangle Park, North Carolina 27711, U.S.A.

SUMMARY

The 'EM' (Expectation-Maximization) algorithm is applied to probit analysis with multiple independent variables and a nonzero response rate. The equations for the maximum likelihood estimators are relatively simple, and converge in all the cases so far examined. An animal bioassay example is used to demonstrate the method. Likelihood ratio tests are used for significance tests in the analysis.

1. Introduction

Probit analysis has been described by Finney (1971) in his classic work. Although Finney describes analyses using multiple independent variables, these analyses require complicated maximization algorithms, such as the Gauss–Newton method, in order to obtain convergence. The case of a natural response rate (nonzero background) can be particularly troublesome. This communication gives an algorithm for fitting a probit curve with multiple independent variables and a natural response rate; this algorithm is computationally simple and converges under a wide range of conditions.

2. The Method

Maximum likelihood estimation was greatly enhanced by the statement of the 'Missing Information Principle' (MIP) by Orchard and Woodbury (1972). This principle provides that the missing information be estimated from the data present, and that the parameters then be reestimated from the 'completed' information. This method was termed the 'EM' algorithm by Dempster, Laird and Rubin (1977), and its general convergence was shown. The 'EM' algorithm is defined as the alternate use of E- and M-steps, where

E-step: estimate the complete-data sufficient statistics from the observed data and previous parameter estimates;

M-step: estimate the parameters from the estimated sufficient statistics.

These steps are now developed for the multiple probit problem.

Multiple probit analysis with a nonzero background can be thought of as the analysis of a mixture of two binomial distributions: a probit distribution with success probability $p = \Phi(\mu)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and a degenerate distribution with success probability $p = 1$. The mixing proportions are $1 - \lambda$ and λ, respectively, where λ is the nonzero background. Furthermore, the probit distribution can be thought of as the dichotomy of a conditionally normally-distributed random

Key words: Maximum likelihood; EM algorithm; Missing information principle; Natural response rate.

variable. Let

$$z_i = \sum_{j=1}^{K} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$
$$= \mu_i + \varepsilon_i, \tag{1}$$

where the x_{ij} are known and fixed, and ε_i is normally distributed with mean zero and unit variance. The value of x_{i1} is commonly taken to be identically 1. If we define y_i as

$$y_i = \begin{cases} 1 & \text{if } z_i > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

then

$$\text{pr}(y_i = 1) = \lambda + (1-\lambda)\Phi(\mu_i). \tag{3}$$

Since a mixed distribution can also be regarded as a missing information problem, the EM algorithm can be used here. All '0' responses, by definition, belong to the probit distribution, whereas the '1' responses can belong to either distribution. The E-step of the EM algorithm consists of two computations. Each '1' response is first apportioned to the respective distributions by the relative estimated probabilities

$$\text{pr}(z_i \in \text{Probit} \mid y_i = 1) = w_i = (1-\lambda)\hat{\mu}_i / \{\lambda + (1-\lambda)\Phi(\hat{\mu}_i)\}, \tag{4}$$

where

$$\hat{\mu}_i = \sum_{j=1}^{K} \hat{\beta}_j x_{ij}.$$

The conditional expected values of the z_i variables are

$$\hat{E}(z_i \mid y_i = 0) = \hat{E}(z_i \mid z_i \leq 0) = \hat{\mu} - \phi(\hat{\mu})/\{1 - \Phi(\hat{\mu})\} \tag{5}$$

and

$$\hat{E}(z_i \mid y_i = 1) = \hat{\mu} + \phi(\hat{\mu})/\Phi(\hat{\mu}),$$

where $\phi(\cdot)$ is the standard normal density.

The M-step then reestimates the β_j parameters using the familiar maximum likelihood equations. If the known matrix of x_{ij} is denoted by $\mathbf{X}$, the diagonal matrix of w_i by $\mathbf{W}$, and the vector of the estimated z_i by $\mathbf{Z}$, we find the estimator of the vector $\boldsymbol{\beta}$ to be

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{Z}), \tag{6}$$

where the elements of the matrices $\mathbf{X}^T\mathbf{W}\mathbf{X}$ and $\mathbf{X}^T\mathbf{W}\mathbf{Z}$ are

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})_{r,s} = \sum_{i=1}^{n} w_i x_{ir} x_{is}$$

and

$$(\mathbf{X}^T\mathbf{W}\mathbf{Z})_r = \sum_{i=1}^{n} w_i x_{ir} z_i.$$

Thus $\hat{\boldsymbol{\beta}}$ is simply a weighted least squares estimator. Further, $\hat{\lambda} = \sum_{i=1}^{n} (1 - w_i) y_i / n$.

In some experiments there are counts of positive and negative controls, which will be denoted as b_1 and b_2, respectively. Incorporating these into the estimate of λ gives

$$\hat{\lambda} = \left\{ \sum_{i=1}^{n} (1 - w_i) y_i + b_1 \right\} \Big/ (n + b_1 + b_2). \tag{7}$$

Table 1
Mortality of mice exposed to NO_2

NO_2 (ppm)	Time (hours)	Number tested	Number dead
1.5	96	120	44
	168	80	37
	336	80	43
	504	60	35
	Controls	340	74
3.5	0.5	100	29
	1	200	53
	2	40	13
	3	200	75
	5	40	23
	7	280	152
	14	80	55
	24	140	98
	48	160	121
	Controls	280	56
7.0	0.5	120	52
	1	120	62
	1.5	120	61
	2	120	86
	Controls	120	27

3. An Example

Larsen *et al.* (1977) presented an experiment that measured mortality in mice in relation to NO_2 exposure. The experiment consisted of exposing mice to NO_2 concentrations ranging from 0.04 to 526.7 mg/m^3 (0.5–28.0 ppm) for time spans ranging from six minutes up to one year. The mice were then challenged with a viable microorganism. The data in Table 1 represent a subset from the original study. Because of natural mortality, a control group was also tested at each NO_2 level.

For this example, a number of probit models were fitted to the data. The most general includes separate background rates, intercepts, and slopes for each NO_2 level, while the simplest model includes only a common background with a constant term plus coefficients for log(NO_2) and log(time). In Table 2, the results are shown, which lead to the likelihood ratio tests given in Table 3.

By using the EM algorithm, this complex problem was handled in a straightforward manner using a general iterative procedure to fit the data. The data are shown in Fig. 1 along with the fitted model (C).

4. Discussion

Application of a multiple probit model to data, especially when a natural response rate is present, has in the past presented computational difficulties. The multiple probit without a natural response rate has recently been available as a standard option in GLIM (Baker and Nelder, 1978). Now, however, the iterative methods presented in this communication allow the use of a probit analysis in rather complex experiments and include allowance for

Table 2
Fits of various models to the mice mortality data

Model	Log likelihood	Number of parameters	Goodness of fit*	df	*P*-value
(A) Separate λ values, intercepts, slopes	−1717.2091	9	14.453	11	.209
(B) Common λ, separate intercepts, slopes	−1717.5688	7	15.173	13	.297
(C) Common λ, intercept, $\ln(NO_2)$, ln(time), $\ln(NO_2) \times \ln$(time),	−1717.7817	5	15.599	15	.409
(D) Common λ, intercept, $\ln(NO_2)$, ln(time)	−1723.7281	4	27.491	16	.036

* Tested using likelihood ratios against the general multinomial model.

Table 3
Likelihood ratio tests for the mice mortality data

Test	Asymptotic chi-square	df	*P*-value
Common λ (A vs B)	.719	2	.698
Separate lines vs linear model with interaction (B vs C)	.426	2	.808
Interaction term (C vs D)	11.893	1	<.001

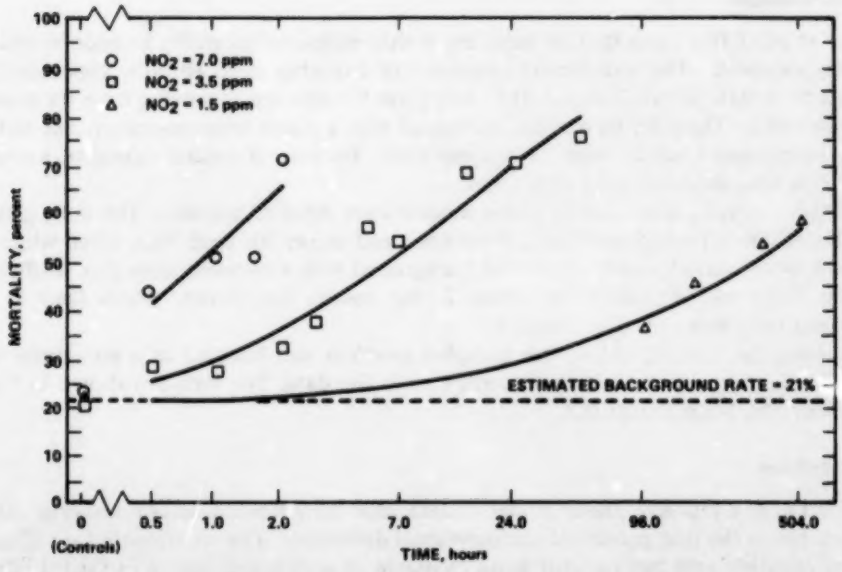


Figure 1. Mortality in mice exposed to NO_2 with fitted model. $P(NO_2, \text{Time}) = .210 + .790\Phi\{-2.94 + .309 \ln(\text{time}) + 1.374 \ln(NO_2) + .138 \ln(\text{time}) \times \ln(NO_2)\}$

a natural response rate. The procedure is intuitively appealing and computationally direct and simple. A FORTRAN computer program has been written to implement the procedure for experiments that either do or do not include background rates. Convergence to maximum likelihood solutions has invariably been rapid in the data sets so far examined. This probit analysis procedure has been applied in a variety of other situations involving a probit-type response variable. We have found the techniques to be effective in all these cases.

Acknowledgements

The authors wish to thank the Editor and referee for their excellent suggestions. The authors are also grateful to Barbara Crabtree for her secretarial help.

Résumé

L'algorithme 'EM' (Expectation-Maximization) est appliqué à l'analyse de probit d'un taux de réponse non nul, à partir de plusieurs variables indépendantes. Les équations donnant les estimateurs de maximum de vraisemblance sont relativement simples, et leurs solutions ont toujours convergé jusqu'à présent. La méthode est illustrée sur un exemple d'essai biologique d'origine animale. Les tests statistiques utilisés dans l'analyse sont des tests de rapports de vraisemblance.

References

Baker, R. J. and Nelder, J. A. (1978). *The GLIM System, Release* 3. Oxford: Numerical Algorithms Group.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **38,** 1–11.

Finney, D. J. (1971). *Probit Analysis,* 3rd ed. Cambridge: Cambridge University Press.

Larsen, R. I., Gardner, D. E. and Coffin, D. L. (1979). An air quality data analysis system for interrelating effects, standards, and needed source reductions: Part 5. NO_2 mortality in mice. *Journal of the Air Pollution Control Association* **39,** 133–137.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* **1,** 697–715.

Received March 1980; *revised June* 1980